(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 07110814 A

(43) Date of publication of application: 25.04.95

(51) Int. Cl G06F 17/27
// G06F 12/00

(21) Application number: 05277337

(22) Date of filing: 12.10.93

(71) Applicant: FUJI XEROX CO LTD

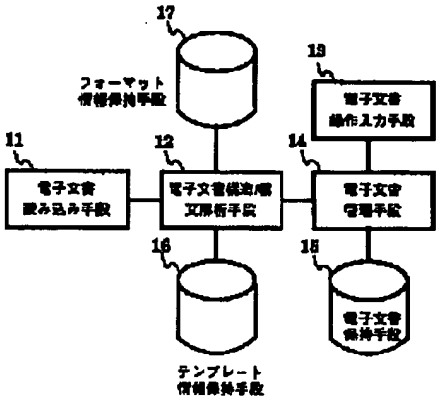
(72) Inventor: NAKATSUYAMA HISASHI
KURAHASHI MASAYUKI

(54) STORAGE DEVICE FOR ELECTRONIC DOCUMENT

(57) Abstract:

PURPOSE: To collectively and automatically process electronic documents stored in the same holder by sorting these documents based on their structure.

CONSTITUTION: This electronic document storage device is constituted of an electronic document reading means 11 for reading an electronic document, an electronic document syntax analyzing means 12 for analyzing the syntax of the read electronic document and an electronic document storing means 15 for storing the electronic document in a hierarchical storage position based on the syntax of the electronic document analyzed by the means 12.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平7－110814

(43)公開日　平成７年(1995)４月25日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/27					
// Ｇ０６Ｆ　12/00	５０５	8944－5B			
		7315－5L	Ｇ０６Ｆ　15/ 20	５５０　Ｅ	

審査請求　未請求　請求項の数１　ＦＤ　（全　８　頁）

(21)出願番号　特願平5－277337

(22)出願日　平成５年(1993)10月12日

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂三丁目３番５号

(72)発明者　中津山　恒
神奈川県横浜市保土ヶ谷区神戸町134番地
横浜ビジネスパークイーストタワー　富士ゼロックス株式会社内

(72)発明者　倉橋　政之
神奈川県海老名市本郷2274　富士ゼロックス株式会社内

(74)代理人　弁理士　加藤　恭介　（外３名）

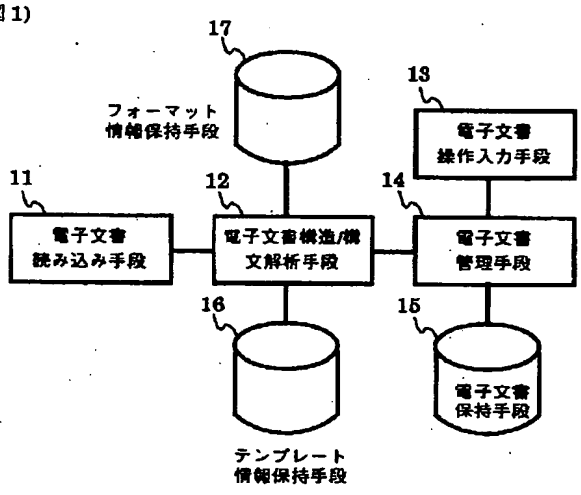
(54)【発明の名称】　電子文書記憶装置

(57)【要約】

【目　的】　文書構造に基づいて電子文書を分類することによって、同一のフォルダに格納されている電子文書を一括して自動処理する。

【構　成】　電子文書記憶装置は、電子文書を読み込む電子文書読み込み手段（１１）と、前記電子文書読み込み手段（１１）により読み込まれた電子文書の構造を解析する電子文書構造解析手段（１２）と、前記電子文書構造解析手段（１２）により解析された電子文書の構造に基づいて階層化されている記憶位置に電子文書を記憶する電子文書記憶手段（１５）とから構成される。

(図1)



1

【特許請求の範囲】
【請求項１】　電子文書を読み込む電子文書読み込み手段と、
前記電子文書読み込み手段により読み込まれた電子文書の構造を解析する電子文書構造解析手段と、
前記電子文書構造解析手段により解析された電子文書の構造に基づいて階層化されている記憶位置に電子文書を記憶する電子文書記憶手段と、
を具備することを特徴とする電子文書記憶装置。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、パーソナルコンピュータ、ワークステーション、あるいはワードプロセッサ等によって作成された電子文書を効率良くファイリングすることができる電子文書記憶装置に関するものである。
【０００２】
【従来の技術】従来、パーソナルコンピュータ、ワークステーション、あるいはワードプロセッサ等によって作成された電子文書を電子文書記憶装置にファイリングする際に先立って、ユーザは、手作業で電子文書に分類付けを行う。そして、この分類に基づいて、電子文書は、電子文書記憶装置等にファイリングされていた。上記従来の電子文書記憶装置は、上記電子文書をファイリングする際に、通常のファイルをキャビネットや書棚に格納する際の手法を模倣していた。たとえば、電子文書記憶装置にキャビネットやバインダーに相当するフォルダ（格納場所）を備え、ユーザは、このフォルダ毎に格納可能な電子文書の種類を定め、この分類規則にしたがって、電子文書のファイリングあるいは検索を行なっていた。
【０００３】上記のように、ユーザは、定められた分類規則を遵守して、電子文書を手作業によって電子文書記憶装置にファイリングするため、作業能率が悪い。そこで、上記欠点を除去するために、ユーザが手作業で分類付けを行なうことなく、分類を自動化する文書管理装置が考案されている。たとえば、特開平２－９６２６８号公報における「文書管理装置」では、電子文書に予め付与されたキーワードを利用して、システムの定められた格納場所に電子文書を分類する方式が記載されている。この文書管理装置では、ディレクトリ毎に格納可能な電子文書の条件が指定される。たとえば、「あるディレクトリには、『データベース』というキーワードが与えられている電子文書だけが格納できる」という情報をシステムが保持する。この条件を利用して、電子文書が自動的に分類される。このような手法を利用すると、分類条件を変更することによって、再分類することが可能になっている。
【０００４】
【発明が解決しようとする課題】特開平２－９６２６８号公報における「文書管理装置」に示された分類の情報

2

としては、キーワードとキーセンテンスをあげている。しかし、これらのキーワードおよびキーセンテンスは、電子文書がもともと備えていた情報ではなく、格納時にユーザが付与するものである。また、従来の電子文書ファイリング装置が取り扱う電子文書は、印刷された文書をスキャナーで読み取ったイメージである。したがって、電子文書ファイリング装置は、文書が構造をもたないので、文書構造に基づく処理の対象にはなり得なかった。上記手法は、構造をもった電子文書として、管理することが可能である。しかし、電子文書をキーワードやキーセンテンスで分類した場合には、同一のフォルダにあるにもかかわらず、電子文書間の文書構造が異なることがある。その電子文書間の文書構造が異なるため、電子文書の一部を取り出すような操作を行なう場合には、フォルダに格納されているすべての電子文書に同じ操作を適用できる保証がなかった。この問題は、電子文書の分類に文書構造が考慮されていないことに起因する。
【０００５】本発明は、以上のような課題を解決するためのもので、文書構造に基づいて電子文書を分類することによって、同一のフォルダに格納されている電子文書を一括して自動処理することができる電子文書記憶装置を提供することを目的とする。
【０００６】
【課題を解決するための手段】前記目的を達成するために、本発明の電子文書記憶装置は、電子文書を読み込む電子文書読み込み手段（図１の１１）と、前記電子文書読み込み手段（１１）により読み込まれた電子文書の構造を解析する電子文書構造解析手段（図１の１２）と、前記電子文書構造解析手段（１２）により解析された電子文書の構造に基づいて階層化されている記憶位置に電子文書を記憶する電子文書記憶手段（図１の１５）とから構成される。
【０００７】
【作　　用】電子文書読み込み手段によって読み込まれた電子文書は、たとえば、予めテンプレート情報保持手段に設定されたテンプレート情報、あるいはフォーマット情報保持手段に格納された電子文書のフォーマット情報を基に、電子文書構造解析手段により構造が解析される。次に、電子文書構造解析手段による電子文書の構造を解析した結果は、たとえば電子文書管理手段に渡される。電子文書管理手段は、電子文書の構造が解析された結果に基づく文書構造に応じた電子文書記憶手段の適切な場所に保存される。したがって、電子文書は、その構造に基づいて自動的に電子文書記憶手段に保存され、構造的に分類された電子文書の一部を上記電子文書記憶手段から読み出すことができる。
【０００８】
【実　施　例】図１は本発明の一実施例を説明するためのブロック構成図である。図１において、電子文書記憶装置は、電子文書を読み込む電子文書読み込み手段１１

と、電子文書読み込み手段１１によって読み込まれた電子文書の構造／構文を解析する電子文書構造／構文解析手段１２と、電子文書の構造に基づいた検索操作を入力する電子文書操作入力手段１３と、当該電子文書操作入力手段１３の入力操作を解釈して電子文書を検索する電子文書管理手段１４と、構造／構文解析の結果に基づく文書構造で保持する電子文書保持手段１５と、電子文書の構造／構文を解析するために、予め保持するテンプレート情報保持手段１６と、異なるフォーマットからなる電子文書をテンプレート処理ができるように変換するための情報を備えたフォーマット情報保持手段１７とから構成される。

【０００９】本実施例では、電子文書構造／構文解析手段１２によって解析された結果に基づき、電子文書を電子文書保持手段１５における複数のフォルダに分類し格納する。予め決められた構造／構文の電子文書は、各フォルダに格納されるようになっており、該当するフォルダがない場合、特別に用意された「その他」のフォルダに格納される。「その他」のフォルダを備えることで、予想されていない電子文書は、「その他」のフォルダに分類された後、そのフォルダに格納される。「その他」のフォルダに分類された電子文書については、後処理として、既存のフォルダに合うように電子文書の構造を変更するか、または、その電子文書の構造に合うフォルダを新規に作成して、そのフォルダに格納するなどの処理を行うことができる。

【００１０】本実施例では、電子文書を分類するために、テンプレート情報を用いる。ＳＧＭＬ（Ｓｔａｎｄａｒｄ　Ｇｅｎｅｒａｌｉｚｅｄ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ，ＩＳＯ８８７９）の文書のように、文書構造があらかじめ定められている電子文書は、タイトル、内容などの文書の構成要素が決まっている。これらの情報をテンプレートとし、入力文書データとマッチングを行う。この場合、テンプレートは、ＳＧＭＬのＤＴＤ（Ｄｏｃｕｍｅｎｔ　Ｔｙｐｅ　Ｄｅｆｉｎｉｔｉｏｎ　以下、本明細書において、単にＤＴＤと記載する）に準じたものとなる。以下は、記事のＤＴＤの例である。

```
＜！ＤＯＣＴＹＰＥ　記事　［
＜！ＥＬＥＭＥＮＴ　記事　　－Ｏ（タイトル，本文）
＞
＜！ＥＬＥＭＥＮＴ　タイトル－Ｏ（＃ＰＣＤＡＴＡ）
＞
＜！ＥＬＥＭＥＮＴ　本文　　－Ｏ（＃ＰＣＤＡＴＡ）
＞
］＞
```

【００１１】このＤＴＤで定められる電子文書は、上記記載から、電子文書のタイプが「記事」であることが判る。そして、「記事」は、タイトルと本文という要素から構成されていることが判る。また、タイトルと本文の内容は、文字列である。このＤＴＤから作られた例を以下に示す。

```
＜記事＞
＜タイトル＞サンプル＜／タイトル＞
＜本文＞ここには本文が書かれています。＜／本文＞
＜／記事＞
```

＜記事＞は、ここから文書構造が記述されることを示し、＜／記事＞で終了していることを示す。

＜タイトル＞と＜／タイトル＞にはさまれた文字列「サンプル」は、要素「タイトル」の内容である。＜本文＞と＜／本文＞にはさまれた文字列「ここには本文が書かれています。」は、要素「本文」の内容である。

【００１２】以下は、論文のＤＴＤの例である。

```
＜！ＤＯＣＴＹＰＥ　論文　［
＜！ＥＬＥＭＥＮＴ　論文　　－Ｏ（タイトル，著者，サマリ，本文，参考文献）＞
＜！ＥＬＥＭＥＮＴ　タイトル－Ｏ（＃ＰＣＤＡＴＡ）
＞
＜！ＥＬＥＭＥＮＴ　著者　　－Ｏ（＃ＰＣＤＡＴＡ）
＞
＜！ＥＬＥＭＥＮＴ　サマリ　－Ｏ（＃ＰＣＤＡＴＡ）
＞
＜！ＥＬＥＭＥＮＴ　本文　　－Ｏ（＃ＰＣＤＡＴＡ）
＞
＜！ＥＬＥＭＥＮＴ　参考文献－Ｏ（＃ＰＣＤＡＴＡ）
＞
］＞
```

【００１３】このＤＴＤで定められる電子文書は、上記記載から、電子文書のタイプが「論文」であることが判る。そして、「論文」は、タイトル、著者、サマリ、本文、参考文献という要素から構成されていることが判る。また、タイトル、著者、サマリ、本文、参考文献の内容は、文字列である。このＤＴＤから作られた例を以下に示す。

```
＜論文＞
＜タイトル＞電子文書構造化について＜／タイトル＞
＜著者＞倉橋政之＜／著者＞
＜サマリ＞サマリが書かれています。＜／サマリ＞
＜本文＞ここには本文が書かれています。＜／本文＞
＜参考文献＞特開平２－９６２６８号公報＜／参考文献＞
＜／論文＞
```

【００１４】図２は本発明の一実施例で、テンプレート情報を模式的に表したものである。ワードプロセッサを用いて作成された電子文書のように、文書構造があらかじめ定められていない電子文書については、各電子文書の種類に応じて、その構成要素（たとえば、タイトル、著者など）を列挙し、その位置、活字サイズ、文字修飾、キーワードについてのヒント情報を持つことにより、構造解析の精度を上げるようになっている。図１に

おいて、電子文書構造／構文解析手段１２は、電子文書の構造／構文を解析し、テンプレート情報とのマッチングを行う。たとえば、ＬＡＴＥＸ（文書処理システムＬＡＴＥＸ，Ｌ．Ｌａｍｐｏｒｔ，アスキー出版局　参照）などの文書フォーマットの場合は、テンプレート情報とその文書フォーマットが使用するスタイル情報との対応表としてフォーマット情報保持手段１７に持つことにより行う。その他の本質的に構造を持たない文書フォーマットの場合は、フォーマット情報保持手段１７に保持された情報をもとに段落の判定を行い、テンプレート情報保持手段１６に格納された、位置、活字サイズ、文字修飾、キーワードについてのヒント情報と、図１に示す電子文書構造／構文解析手段１２における解析の結果を元に文書構造を判断する。

【００１５】以上により、論文２０の場合は、一般的にタイトル２１、著者２２、サマリ２３、本文２４、参考文献２５等から構成される。そして、たとえば、タイトル２１には、その位置、活字サイズ、タイトル文字の修飾、キーワード等が記述されている。以下、著者２２ないし参考文献２５についても同様である。図３は本発明の一実施例で、記事を表すテンプレート情報の一例を示す図である。図３において、記事３０は、タイトル３１と本文３２とから構成される。図４は本発明の一実施例で、カタログを表すテンプレート情報の一例を示す図である。図４において、カタログ４０は、タイトル４１、画像４２、テキスト（内容の説明）４３、商品番号４４、価格４５から構成される。図５は本発明の一実施例で、マニュアルを表すテンプレート情報の一例を示す図である。図５において、マニュアル５０は、タイトル５１、使い方のサマリ５２、目次５３、詳細な内容５４、索引５５から構成される。

【００１６】そして、電子文書構造／構文解析手段１２は、たとえば図２ないし図５に示すテンプレート情報によって電子文書を解析し、その結果に基づいて電子文書を電子文書保持手段１５における「論文」、「記事」、「カタログ」、「マニュアル」というフォルダにそれぞれ格納する。電子文書構造／構文解析手段１２は、もし、上記分類にあてはまらない場合、その電子文書を「その他」というフォルダに格納する。図６は本発明の一実施例で、フォルダによる分類を模式的に表したものである。図６において、「電子文書」６１というフォルダの中に、「論文」６２、「記事」６３、「その他」６４というフォルダが入り、それぞれのフォルダの中には、そのフォルダに分類された電子文書が入っている。

【００１７】上記実施例において、テンプレート情報の定義方法により、分類に階層を設定することができる。たとえば、図２に示すテンプレート情報の場合、「論文」は、“著者”、“サマリ”、“本文”、“参考文献”をまとめて、“本文”と設定することにより、「記事」としても分類が可能である。この場合、「記事」という大きい分類に対応するフォルダの中に、より詳細な分類、すなわち、「論文」に対応するフォルダを設けることにより、フォルダに階層を設ける。

【００１８】図７は、本発明の一実施例で、階層化されたフォルダによる分類を模式的に表したものである。図７に示す「記事」は、その中に、「論文」というフォルダが入っている例である。図７に示す例では、ユーザが論文である電子文書７１を「論文」７６というフォルダで取り出すこともできるし、より大きな分類である「記事」７２というフォルダでも取り出すことができる。本実施例は、テンプレート情報が階層化されており、下位の階層のテンプレート情報ほど、より詳細な構造を表現することになる。

【００１９】図８は本発明の実施例におけるテンプレートの階層情報を模式的に表した図である。このテンプレートの階層情報は、たとえば電子文書８１、記事８２、論文８３、およびその他８４から構成され、図１に示すテンプレート情報保持手段１６にテンプレート情報と共に格納されている。そして、図１に示す電子文書構造／構文解析手段１２は、このテンプレートの階層情報を基に上位のテンプレート情報から順次マッチングを行う。

【００２０】図９は本発明の実施例におけるテンプレートのクラス階層情報を模式的に表した図である。クラス階層情報は、たとえば電子文書クラス９１と、記事文書クラス９２と、論文文書クラス９３とから構成される。そして、電子文書クラス９１には、可能なオペレーションとして、「全文の取り出し」がある。論文文書クラス９３には、可能なオペレーションとして、たとえば、「タイトルの取り出し」、「著者の取り出し」、「サマリの取り出し」、「論文本文の取り出し」、「参考文献の取り出し」がある。図１に示す電子文書構造／構文解析手段１２は、電子文書の構造／構文を解析した結果に基づき、電子文書を予め設定しておいた電子文書クラスに分ける。

【００２１】電子文書保持手段１５には、これらの電子文書クラスが定める構造に応じた形で文書が保存される。本実施例では、文書全体を一つの単位として保存するのではなく、テンプレートでマッチングした電子文書の構造単位で保存する。たとえば、図２に示す「論文」のテンプレート情報を用いて、マッチングした場合は、「タイトル」、「著者」、「サマリ」、「本文」、「参考文献」という単位で電子文書保持手段１５に格納される。図９に示すクラス階層情報は、テンプレート情報保持手段１６にテンプレート情報と共に格納される。図１に示す電子文書構造／構文解析手段１２は、文書の分類にあたって、上位のクラスから順次マッチングを行う。電子文書は、マッチしたクラスのうち、もっとも下位のクラスに分類される。これは、下位のクラスほど構造が詳細化されており、細かい処理に適しているからである。以上のような形で電子文書が電子文書保持手段１５

に格納されているため、電子文書操作入力手段１３は、クラス階層情報を基にして、全文を取り出したり、タイトルを取り出したり、あるいは参考文献を取り出すことができる。そして、電子文書の取り出し方は、一つの論文の参考文献を取り出したり、あるいは全ての論文に付けられている全参考文献を取り出すようなことも可能である。

【００２２】図１０は本発明の一実施例で、電子文書を分類して電子文書保持手段に格納する際のフローチャートを示す。図１０において、電子文書構造／構文解析手段１２は、テンプレート情報保持手段１６から、ユーザの所望するように分類されているテンプレート情報を入力する（ステップ１０１）。電子文書構造／構文解析手段１２は、フォーマット情報保持手段１７から、電子文書の構造を解析するためのフォーマット情報を入力する（ステップ１０２）。電子文書構造／構文解析手段１２は、電子文書読み込み手段１１から、電子文書保持手段１５に分類付けをして格納する電子文書データを入力する（ステップ１０３）。

【００２３】電子文書構造／構文解析手段１２は、入力された電子文書データが終わりか否かを調べる（ステップ１０４）。電子文書構造／構文解析手段１２は、入力された電子文書データが終わりであると判断した場合、処理を終了させる。電子文書構造／構文解析手段１２は、テンプレートによってマッチング処理を行なう（ステップ１０５）。電子文書構造／構文解析手段１２は、ステップ１０５のマッチング処理による分類結果に応じて、前記電子文書を電子文書保持手段１５に格納処理する。また、電子文書構造／構文解析手段１２は、次の電子文書データを入力するために処理をステップ１０３に戻す（ステップ１０６）。

【００２４】次に、フォーマットの異なる電子文書が図１に示す電子文書読み込み手段１１によって読み込まれた場合を説明する。フォーマットの異なる電子文書が読み込まれた場合、そのままでは、テンプレート情報によるマッチング処理を行なうことができない。たとえば、ＪＩＳ文書とシフトＪＩＳ文書とでは、制御文字が異なるため、前述のような処理を行なうことができない。そこで、電子文書構造／構文解析手段１２は、フォーマット情報保持手段１７に格納されているフォーマット対応表に基づき、前記テンプレート情報が利用できる状態に変換する。たとえば、フォーマット情報保持手段１７には、ＪＩＳ文書とシフトＪＩＳ文書とにおけるフォーマットの対応表を持ち、この表に基づいて一方の文書に変換する。その後、前記電子文書は、前述のような処理を行なうことで、自動的に分類して電子文書保持手段１５に格納する。

【００２５】

【発明の効果】本発明によれば、電子文書構造／構文解析手段によって、電子文書の構造を解析し、その結果に基づいて自動的に分類した後、電子文書保持手段に格納するため、電子文書にキーワードあるいはキーセンテンスのような分類を付与する必要がなくなる。また、本発明によれば、電子文書を階層化した状態で、電子文書保持手段に格納しているため、電子文書保持手段における同一箇所に分類されたすべての文書に対し、同一の操作が適用できる。上記のように分類して電子文書保持手段に格納されている電子文書は、電子文書構造／構文解析手段によって、電子文書の構造の一部を取り出すことができる。

【図面の簡単な説明】

【図１】　本発明の一実施例を説明するためのブロック構成図である。

【図２】　本発明の一実施例で、テンプレート情報を模式的に表したものである。

【図３】　本発明の一実施例で、記事を表すテンプレート情報の一例を示す図である。

【図４】　本発明の一実施例で、カタログを表すテンプレート情報の一例を示す図である。

【図５】　本発明の一実施例で、マニュアルを表すテンプレート情報の一例を示す図である。

【図６】　本発明の一実施例で、フォルダによる分類を模式的に表したものである。

【図７】　本発明の一実施例で、階層化されたフォルダによる分類を模式的に表したものである。

【図８】　本発明の実施例におけるテンプレートの階層情報を模式的に表した図である。

【図９】　本発明の実施例におけるテンプレートのクラス階層情報を模式的に表した図である。

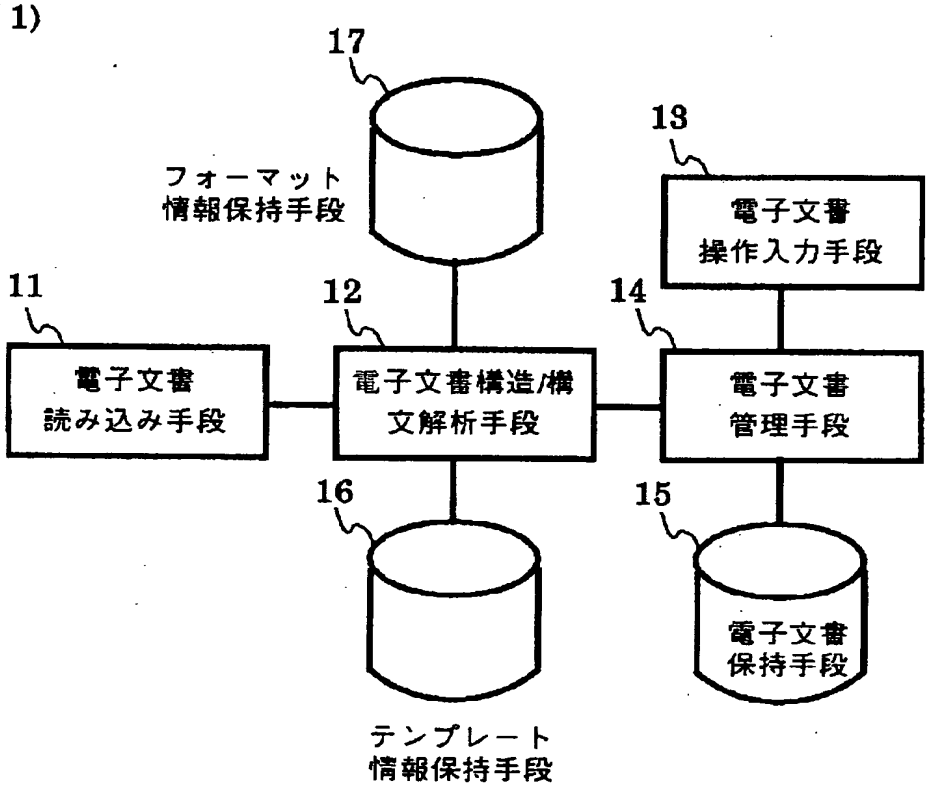
【図１０】　本発明の一実施例で、電子文書を分類して電子文書保持手段に格納する際のフローチャートを示す。

【符号の説明】

１１・・・電子文書読み込み手段
１２・・・電子文書構造／構文解析手段
１３・・・電子文書操作入力手段
１４・・・電子文書管理手段
１５・・・電子文書保持手段
１６・・・テンプレート情報保持手段
１７・・・フォーマット情報保持手段

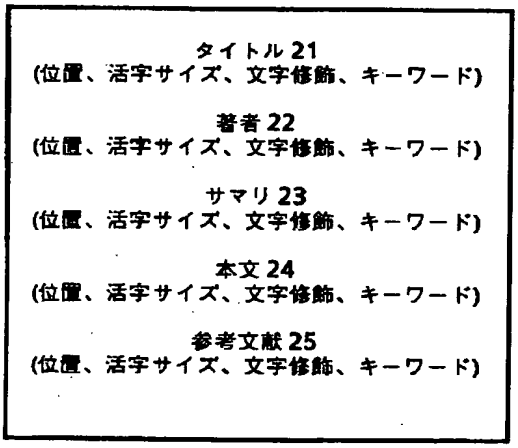
【図１】

(図 1)



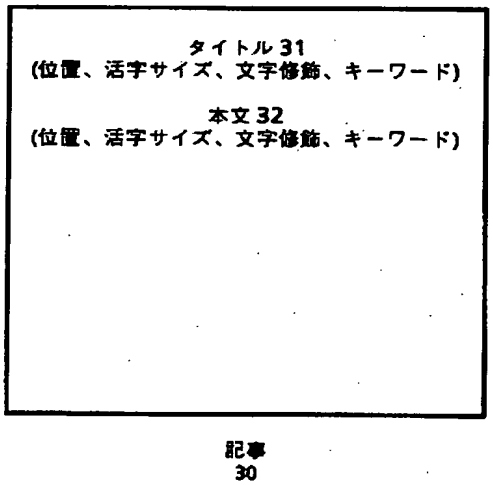
【図２】

(図 2)



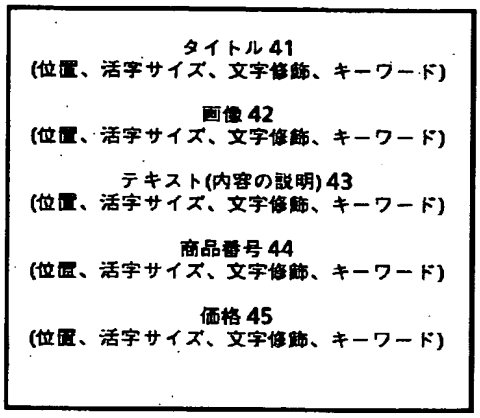
【図３】

(図 3)



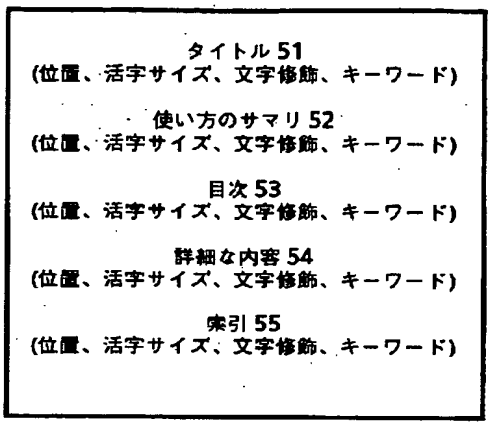
【図４】

(図 4)



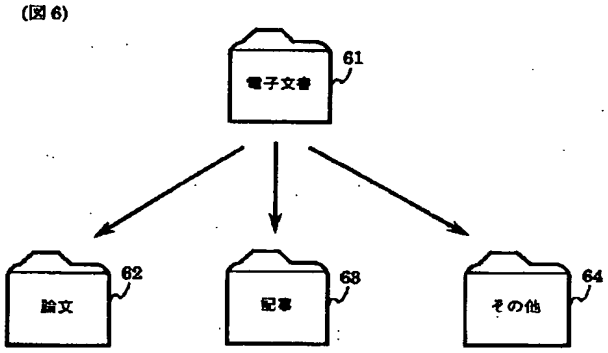
【図５】

(図 5)



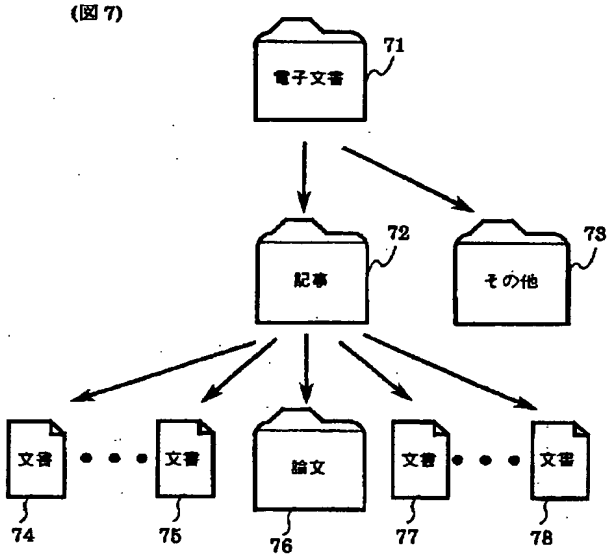
【図６】

(図 6)



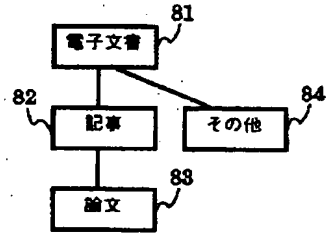
【図７】

(図 7)



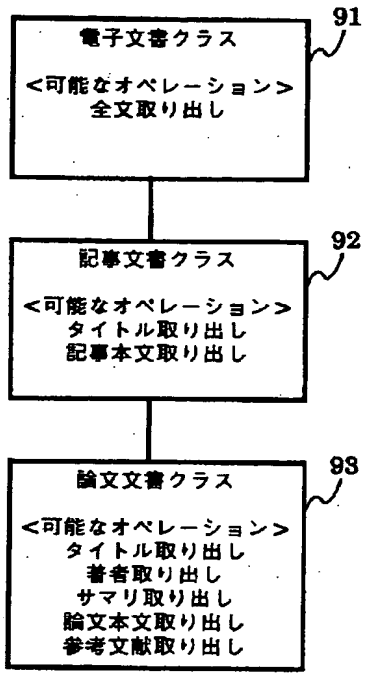
【図８】

(図 8)



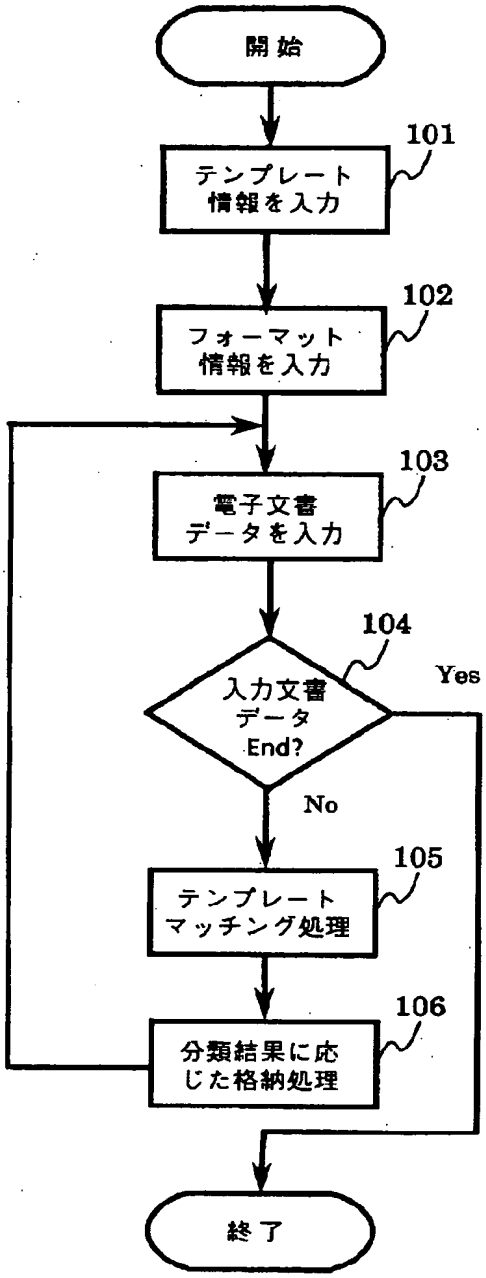
【図９】

(図9)


電子文書クラス
<可能なオペレーション>
全文取り出し
91
記事文書クラス
<可能なオペレーション>
タイトル取り出し
記事本文取り出し
92
論文文書クラス
<可能なオペレーション>
タイトル取り出し
著者取り出し
サマリ取り出し
論文本文取り出し
参考文献取り出し
93

【図１０】

(図10)


開始
テンプレート情報を入力
101
フォーマット情報を入力
102
電子文書データを入力
103
入力文書データEnd?
104
Yes
No
テンプレートマッチング処理
105
分類結果に応じた格納処理
106
終了